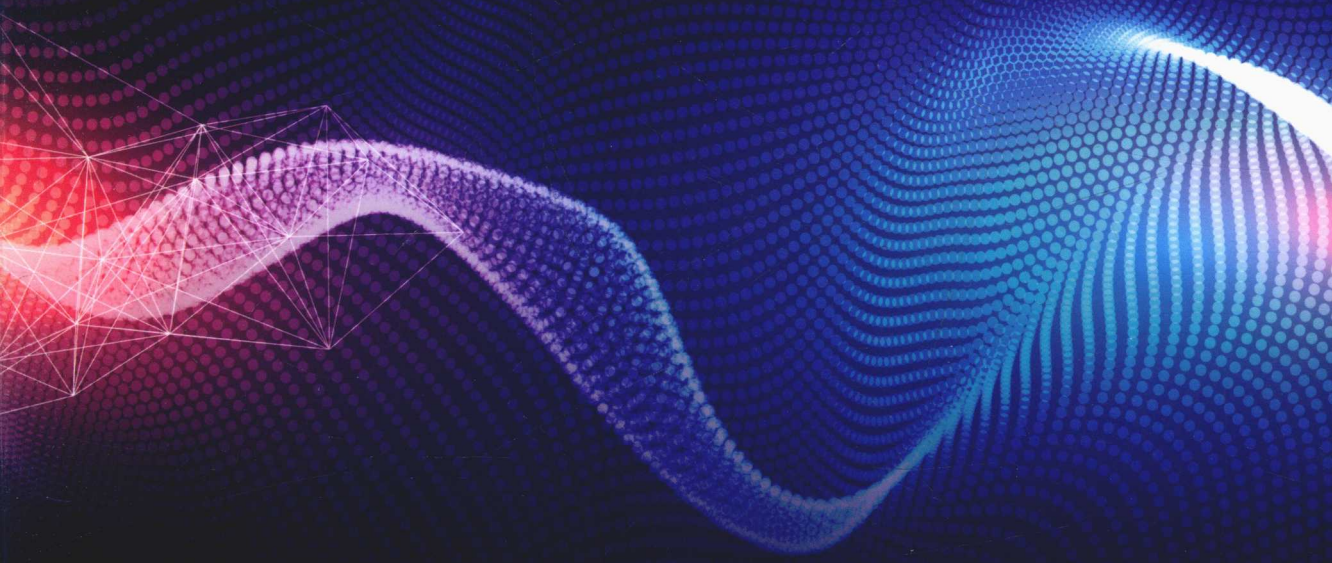
HZ BOOKS
华章IT

资深大数据技术专家/《Hadoop技术内幕》系列图书作者撰写

从数据收集、数据存储、资源管理与服务协调、计算引擎、数据分析5个层次系统、深度剖析大数据技术体系

大数据技术丛书



Big Data Technical System
Principle, Architecture and Practice

大数据技术体系详解
原理、架构与实践

董西成 ◎ 著

机械工业出版社
China Machine Press

技术丛书

Big Data Technical System

Principle, Architecture and Practice

大数据技术体系详解

原理、架构与实践

董西成◎著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据技术体系详解：原理、架构与实战 / 董西成著 . —北京：机械工业出版社，2018.1
（大数据技术丛书）

ISBN 978-7-111-59072-9

I. 大… II. 董… III. 数据处理 IV. TP274

中国版本图书馆 CIP 数据核字（2018）第 023800 号

大数据技术体系详解：原理、架构与实战

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：何欣阳 责任校对：李秋荣
印 刷：北京市荣盛彩色印刷有限公司 版 次：2018 年 3 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：23.5
书 号：ISBN 978-7-111-59072-9 定 价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

华章IT
HZBOOKS | Information Technology

Preface 前　言

为什么要写这本书

随着大数据技术的普及，它已经被广泛应用于互联网、电信、金融、工业制造等诸多行业。据相关报告统计，大数据人才需求呈井喷态势，越来越多的程序员开始学习大数据技术，这使得它已经成为程序员所需的基本技能。

为了满足大数据人才市场需求，越来越多的大数据技术书籍不断面世，包括《Hadoop 权威指南》《Hadoop 实战》等。尽管如此，面向初、中级学者，能够系统化、体系化介绍大数据技术的基础书籍并不多见。笔者曾接触过大量大数据初学者，他们一直渴望能有一本简单且易于理解的教科书式的大数据书籍出现。为了满足这些读者的需求，笔者根据自己多年的数据项目和培训经验，继《Hadoop 技术内幕》书籍之后，于两年前开始尝试编写一本浅显易读的大数据基础书籍。

相比于现有的大数据基础书籍，本书具有三大特色：①系统性：深度剖析大数据技术体系的六层架构；②技术性：详尽介绍 Hadoop 和 Spark 等主流大数据技术；③实用性：理论与实践相结合，探讨常见的大数据问题。本书尝试以“数据生命周期”为线索，按照分层结构逐步介绍大数据技术体系，涉及数据收集、数据存储、资源管理和服务协调、计算引擎及数据分析五层技术架构，由点及面，最终通过综合案例将这些技术串接在一起。

读者对象

（1）大数据应用开发人员

本书用了相当大的篇幅介绍各个大数据系统的适用场景和使用方式，能够很好地帮助大数据应用开发工程师设计出满足要求的程序。

（2）大数据讲师和学员

本书按照大数据五层架构，即数据收集→数据存储→资源管理与服务协调→计算引擎→数据分析，完整介绍了整个大数据技术体系，非常易于理解，此外，每节包含大量代码示例和思考题目，非常适合大数据教学。

（3）大数据运维工程师

对于一名合格的大数据运维工程师而言，适当地了解大数据系统的应用场景、设计原理和架构是十分有帮助的，这不仅有助于我们更快地排除各种可能的大数据系统故障，也能够让运维人员与研发人员更有效地进行沟通。本书可以有效地帮助运维工程师全面理解当下主流的大数据技术体系。

（4）开源软件爱好者

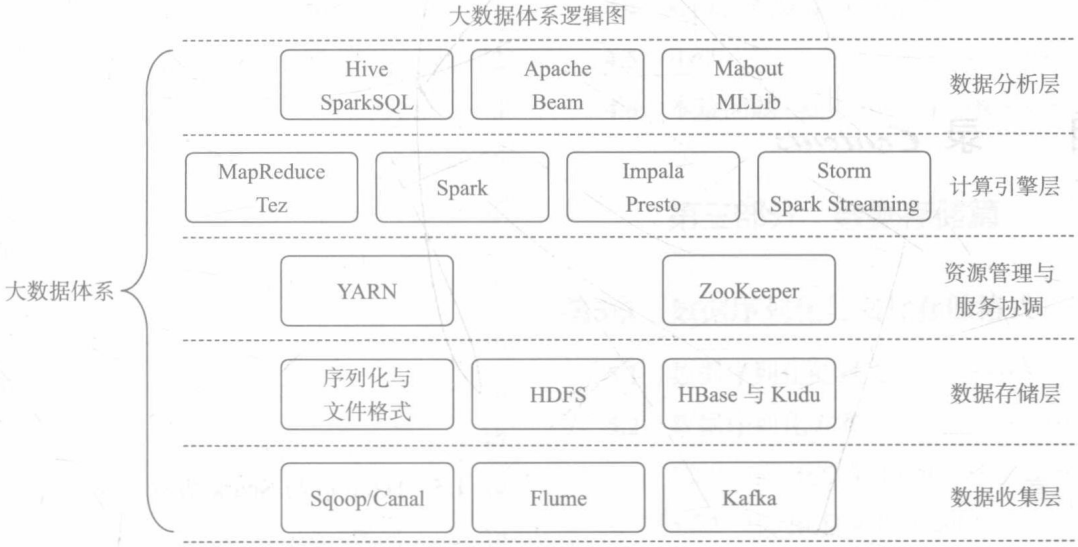
开源大数据系统（比如 Hadoop 和 Spark）是开源软件中的佼佼者，它们在实现的过程中吸收了大量开源领域的优秀思想，同时也有很多值得学习的创新。通过阅读本书，这部分读者不仅能领略到开源软件的优秀思想，还可以学习如何构建一套完整的技术生态。

如何阅读本书

本书以数据在大数据系统中的生命周期为线索，介绍以 Hadoop 与 Spark 为主的开源大数据技术栈。本书内容组织方式如下。

- 第一部分：主要介绍大数据体系架构，以及 Google 和 Hadoop 技术栈，让读者从高层次上对大数据技术有一定了解。
- 第二部分：介绍大数据分析相关技术，主要涉及关系型数据收集工具 Sqoop 与 Canel、非关系型数据收集系统 Flume，以及分布式消息队列 Kafka。
- 第三部分：介绍大数据存储相关技术，涉及数据存储格式、分布式文件系统及分布式数据库三部分。
- 第四部分：介绍资源管理和服务协调相关技术，涉及资源管理和调度系统 YARN，以及资源协调系统 ZooKeeper。
- 第五部分：介绍计算引擎相关技术，包括批处理、交互式处理，以及流式实时处理三类引擎，内容涉及 MapReduce、Spark、Impala/Presto、Storm 等常用技术。
- 第六部分：介绍数据分析相关技术，涉及基于数据分析的语言 HQL 与 SQL、大数据统一编程模型及机器学习库等。

大数据体系的逻辑也是本书的逻辑，故这里给出大数据体系逻辑图。



勘误和支持

由于笔者的水平有限，编写时间仓促，书中难免会出现一些错误或者不准确的地方，恳请读者批评指正。为此，笔者特意创建了一个在线支持与应急方案的站点 http://hadoop123.com 和微信公众号 hadoop-123。你可以将书中的错误发布在 Bug 勘误表页面。如果你遇到任何问题，也可以访问 Q&A 页面，我将尽量在线上为你提供最满意的解答。如果你有更多宝贵的意见，也欢迎发送邮件至邮箱 dongxicheng@yahoo.com，期待能够得到你们的真挚反馈。

获取源代码实例

本书各节的源代码实例可从网站 http://hadoop123.com 或微信公众号 hadoop-123 中获取。

致谢

感谢我的导师廖华明副研究员，是她引我进入大数据世界。

感谢机械工业出版社华章公司的孙海亮编辑对本书的校订，他的鼓励和帮助使我顺利完成了本书的编写工作。

最后感谢我的父母，感谢他们的养育之恩，感谢兄长的鼓励和支持，感谢他们时时刻刻给我以信心和力量！

谨以此书献给我最亲爱的家人，以及众多热爱大数据技术的朋友们！

董西成

目　录 Contents

前　言

第一部分　概述篇

第1章　企业级大数据技术体系概述 …… 2

1.1　大数据系统产生背景及应用场景 … 2

1.1.1　产生背景 …… 2

1.1.2　常见大数据应用场景 …… 3

1.2　企业级大数据技术框架 …… 5

1.2.1　数据收集层 …… 6

1.2.2　数据存储层 …… 7

1.2.3　资源管理与服务协调层 …… 7

1.2.4　计算引擎层 …… 8

1.2.5　数据分析层 …… 9

1.2.6　数据可视化层 …… 9

1.3　企业级大数据技术实现方案 …… 9

1.3.1　Google 大数据技术栈 …… 10

1.3.2　Hadoop 与 Spark 开源大数据技术栈 …… 12

1.4　大数据架构：Lambda Architecture …… 15

1.5　Hadoop 与 Spark 版本选择及安装部署 …… 16

1.5.1　Hadoop 与 Spark 版本选择 …… 16

1.5.2　Hadoop 与 Spark 安装部署 …… 17

1.6　小结 …… 18

1.7　本章问题 …… 18

第二部分　数据收集篇

第2章　关系型数据的收集 …… 20

2.1　Sqoop 概述 …… 20

2.1.1　设计动机 …… 20

2.1.2　Sqoop 基本思想及特点 …… 21

2.2　Sqoop 基本架构 …… 21

2.2.1　Sqoop1 基本架构 …… 22

2.2.2　Sqoop2 基本架构 …… 23

2.2.3　Sqoop1 与 Sqoop2 对比 …… 24

2.3　Sqoop 使用方式 …… 25

2.3.1　Sqoop1 使用方式 …… 25

2.3.2　Sqoop2 使用方式 …… 28

2.4　数据增量收集 CDC …… 31

2.4.1 CDC 动机与应用场景……31
2.4.2 CDC 开源实现 Canal……32
2.4.3 多机房数据同步系统 Otter……33
2.5 小结……35
2.6 本章问题……35

第3章 非关系型数据的收集……36
3.1 概述……36
3.1.1 Flume 设计动机……36
3.1.2 Flume 基本思想及特点……37
3.2 Flume NG 基本架构……38
3.2.1 Flume NG 基本架构……38
3.2.2 Flume NG 高级组件……41
3.3 Flume NG 数据流拓扑构建方法……42
3.3.1 如何构建数据流拓扑……42
3.3.2 数据流拓扑实例剖析……46
3.4 小结……50
3.5 本章问题……50

第4章 分布式消息队列Kafka……51
4.1 概述……51
4.1.1 Kafka 设计动机……51
4.1.2 Kafka 特点……53
4.2 Kafka 设计架构……53
4.2.1 Kafka 基本架构……54
4.2.2 Kafka 各组件详解……54
4.2.3 Kafka 关键技术点……58
4.3 Kafka 程序设计……60
4.3.1 Producer 程序设计……61
4.3.2 Consumer 程序设计……63
4.3.3 开源 Producer 与 Consumer 实现……65
4.4 Kafka 典型应用场景……65
4.5 小结……67
4.6 本章问题……67

第三部分 数据存储篇

第5章 数据序列化与文件存储格式……70
5.1 数据序列化的意义……70
5.2 数据序列化方案……72
5.2.1 序列化框架 Thrift……72
5.2.2 序列化框架 Protobuf……74
5.2.3 序列化框架 Avro……76
5.2.4 序列化框架对比……78
5.3 文件存储格式剖析……79
5.3.1 行存储与列存储……79
5.3.2 行式存储格式……80
5.3.3 列式存储格式 ORC、Parquet 与 CarbonData……82
5.4 小结……88
5.5 本章问题……89

第6章 分布式文件系统……90
6.1 背景……90
6.2 文件级别和块级别的分布式文件系统……91
6.2.1 文件级别的分布式系统……91
6.2.2 块级别的分布式系统……92
6.3 HDFS 基本架构……93
6.4 HDFS 关键技术……94
6.4.1 容错性设计……95
6.4.2 副本放置策略……95

6.4.3 异构存储介质 …… 96
6.4.4 集中式缓存管理 …… 97
6.5 HDFS 访问方式 …… 98
6.5.1 HDFS shell …… 98
6.5.2 HDFS API …… 100
6.5.3 数据收集组件 …… 101
6.5.4 计算引擎 …… 102
6.6 小结 …… 102
6.7 本章问题 …… 103

第7章 分布式结构化存储系统 …… 104

7.1 背景 …… 104
7.2 HBase 数据模型 …… 105
7.2.1 逻辑数据模型 …… 105
7.2.2 物理数据存储 …… 107
7.3 HBase 基本架构 …… 108
7.3.1 HBase 基本架构 …… 108
7.3.2 HBase 内部原理 …… 110
7.4 HBase 访问方式 …… 114
7.4.1 HBase shell …… 114
7.4.2 HBase API …… 116
7.4.3 数据收集组件 …… 118
7.4.4 计算引擎 …… 119
7.4.5 Apache Phoenix …… 119
7.5 HBase 应用案例 …… 120
7.5.1 社交关系数据存储 …… 120
7.5.2 时间序列数据库 OpenTSDB … 122
7.6 分布式列式存储系统 Kudu …… 125
7.6.1 Kudu 基本特点 …… 125
7.6.2 Kudu 数据模型与架构 …… 126
7.6.3 HBase 与 Kudu 对比 …… 126
7.7 小结 …… 127
7.8 本章问题 …… 127

第四部分 分布式协调与资源管理篇

第8章 分布式协调服务ZooKeeper … 130

8.1 分布式协调服务的存在意义 …… 130
8.1.1 leader 选举 …… 130
8.1.2 负载均衡 …… 131
8.2 ZooKeeper 数据模型 …… 132
8.3 ZooKeeper 基本架构 …… 133
8.4 ZooKeeper 程序设计 …… 134
8.4.1 ZooKeeper API …… 135
8.4.2 Apache Curator …… 139
8.5 ZooKeeper 应用案例 …… 142
8.5.1 leader 选举 …… 142
8.5.2 分布式队列 …… 143
8.5.3 负载均衡 …… 143
8.6 小结 …… 144
8.7 本章问题 …… 145

第9章 资源管理与调度系统YARN … 146

9.1 YARN 产生背景 …… 146
9.1.1 MRv1 局限性 …… 146
9.1.2 YARN 设计动机 …… 147
9.2 YARN 设计思想 …… 148
9.3 YARN 的基本架构与原理 …… 149
9.3.1 YARN 基本架构 …… 149
9.3.2 YARN 高可用 …… 152
9.3.3 YARN 工作流程 …… 153

9.4 YARN 资源调度器 …… 155
9.4.1 层级队列管理机制 …… 155
9.4.2 多租户资源调度器产生背景 …… 156
9.4.3 Capacity/Fair Scheduler …… 157
9.4.4 基于节点标签的调度 …… 160
9.4.5 资源抢占模型 …… 163
9.5 YARN 资源隔离 …… 164
9.6 以 YARN 为核心的生态系统 …… 165
9.7 资源管理系统 Mesos …… 167
9.7.1 Mesos 基本架构 …… 167
9.7.2 Mesos 资源分配策略 …… 169
9.7.3 Mesos 与 YARN 对比 …… 170
9.8 资源管理系统架构演化 …… 170
9.8.1 集中式架构 …… 171
9.8.2 双层调度架构 …… 171
9.8.3 共享状态架构 …… 172
9.9 小结 …… 173
9.10 本章问题 …… 173

第五部分 大数据计算引擎篇

第10章 批处理引擎MapReduce …… 176
10.1 概述 …… 176
10.1.1 MapReduce 产生背景 …… 176
10.1.2 MapReduce 设计目标 …… 177
10.2 MapReduce 编程模型 …… 178
10.2.1 编程思想 …… 178
10.2.2 MapReduce 编程组件 …… 179
10.3 MapReduce 程序设计 …… 187
10.3.1 MapReduce 程序设计基础 …… 187
10.3.2 MapReduce 程序设计进阶 …… 194
10.3.3 Hadoop Streaming …… 198
10.4 MapReduce 内部原理 …… 204
10.4.1 MapReduce 作业生命周期 …… 204
10.4.2 MapTask 与 ReduceTask …… 206
10.4.3 MapReduce 关键技术 …… 209
10.5 MapReduce 应用实例 …… 211
10.6 小结 …… 213
10.7 本章问题 …… 213

第11章 DAG计算引擎Spark …… 215
11.1 概述 …… 215
11.1.1 Spark 产生背景 …… 215
11.1.2 Spark 主要特点 …… 217
11.2 Spark 编程模型 …… 218
11.2.1 Spark 核心概念 …… 218
11.2.2 Spark 程序基本框架 …… 220
11.2.3 Spark 编程接口 …… 221
11.3 Spark 运行模式 …… 227
11.3.1 Standalone 模式 …… 229
11.3.2 YARN 模式 …… 230
11.3.3 Spark Shell …… 232
11.4 Spark 程序设计实例 …… 232
11.4.1 构建倒排索引 …… 232
11.4.2 SQL GroupBy 实现 …… 234
11.4.3 应用程序提交 …… 235
11.5 Spark 内部原理 …… 236

11.5.1 Spark 作业生命周期……237
11.5.2 Spark Shuffle……241
11.6 DataFrame、Dataset 与 SQL ……247
11.6.1 DataFrame/Dataset 与 SQL 的关系……248
11.6.2 DataFrame/Dataset 程序设计……249
11.6.3 DataFrame/Dataset 程序实例……254
11.7 Spark 生态系统……257
11.8 小结……257
11.9 本章问题……258

第12章 交互式计算引擎……261

12.1 概述……261
12.1.1 产生背景……261
12.1.2 交互式查询引擎分类……262
12.1.3 常见的开源实现……263
12.2 ROLAP……263
12.2.1 Impala……263
12.2.2 Presto……267
12.2.3 Impala 与 Presto 对比……271
12.3 MOLAP……271
12.3.1 Druid 简介……271
12.3.2 Kylin 简介……272
12.3.3 Druid 与 Kylin 对比……274
12.4 小结……274
12.5 本章问题……274

第13章 流式实时计算引擎……276

13.1 概述……276
13.1.1 产生背景……276
13.1.2 常见的开源实现……278
13.2 Storm 基础与实战……278
13.2.1 Storm 概念与架构……279
13.2.2 Storm 程序设计实例……282
13.2.3 Storm 内部原理……285
13.3 Spark Streaming 基础与实战……290
13.3.1 概念与架构……290
13.3.2 程序设计基础……291
13.3.3 编程实例详解……298
13.3.4 容错性讨论……300
13.4 流式计算引擎对比……303
13.5 小结……304
13.6 本章问题……304

第六部分 数据分析篇

第14章 数据分析语言HQL与SQL…308

14.1 概述……308
14.1.1 背景……308
14.1.2 SQL On Hadoop……309
14.2 Hive 架构……309
14.2.1 Hive 基本架构……310
14.2.2 Hive 查询引擎……311
14.3 Spark SQL 架构……312
14.3.1 Spark SQL 基本架构……312
14.3.2 Spark SQL 与 Hive 对比…313
14.4 HQL……314
14.4.1 HQL 基本语法……314
14.4.2 HQL 应用实例……320

14.5 小结 …… 322
14.6 本章问题 …… 322

第15章 大数据统一编程模型 …… 325

15.1 产生背景 …… 325
15.2 Apache Beam 基本构成 …… 327
15.2.1 Beam SDK …… 327
15.2.2 Beam Runner …… 328
15.3 Apache Beam 编程模型 …… 329
15.3.1 构建 Pipeline …… 330
15.3.2 创建 PCollection …… 331
15.3.3 使用 Transform …… 334
15.3.4 side input 与 side output …… 340
15.4 Apache Beam 流式计算模型 …… 341
15.4.1 window 简述 …… 342
15.4.2 watermark、trigger 与 accumulation …… 344
15.5 Apache Beam 编程实例 …… 346
15.5.1 WordCount …… 346
15.5.2 移动游戏用户行为分析 …… 348
15.6 小结 …… 350
15.7 本章问题 …… 350

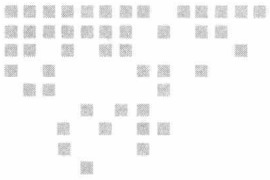
第16章 大数据机器学习库 …… 351

16.1 机器学习库简介 …… 351
16.2 MLLib 机器学习库 …… 354
16.2.1 Pipeline …… 355
16.2.2 特征工程 …… 357
16.2.3 机器学习算法 …… 360
16.3 小结 …… 361
16.4 本章问题 …… 361

第一部分 Part 1

概 述 篇

■ 第 1 章　企业级大数据技术体系概述



Chapter 1 第 1 章

企业级大数据技术体系概述

随着机构和企业积累的数据越来越多，大数据价值逐步体现出来。2015 年国务院向社会公布了《促进大数据发展行动纲要》(以下简称《纲要》)，正式将大数据提升为国家级战略。《纲要》明确提出了大数据的基本概念：**大数据是以容量大、类型多、存取速度快、应用价值高为主要特征的数据集合，正快速发展为对数量巨大、来源分散、格式多样的数据进行采集、存储和关联分析，从中发现新知识、创造新价值、提升新能力的新一代信息技术和服务业态。**《纲要》提到大数据在推动经济转型发展，重塑国家竞争优势，以及提升政府治理能力等方面具有重要的意义，提出在信用、交通、医疗、卫生、金融、气象等众多领域发展大数据。

为了确保大数据思想顺利落地，在各个行业开花结果，需要掌握和利用大数据技术。本书正是从技术角度探讨了如何利用开源技术构建大数据解决方案，从而真正为政府和企业带来实用价值。

1.1 大数据系统产生背景及应用场景

1.1.1 产生背景

大数据技术直接源于互联网行业。随着互联网的蓬勃发展，用户量和数据量越来越多，逐步形成了大数据，这成为大数据技术的基础。根据有关技术报告知道，国内百度、腾讯和阿里巴巴等公司数据规模如下：

- 2013 年百度相关技术报告称，百度数据总量接近 1000PB，网页的数量大是几千亿

个，每年更新几十亿个，每天查询次数几十亿次。

- 2013 年腾讯相关技术报告称，腾讯约有 8 亿用户，4 亿移动用户，总存储数据量经压缩处理以后在 100PB 左右，日新增 200TB 到 300TB，月增加 10% 的数据量。
- 2013 年阿里巴巴相关技术报告称，总体数据量为 100PB，每天的活跃数据量已经超过 50TB，共有 4 亿条产品信息和 2 亿多名注册用户，每天访问超过 4000 万人次。

为了采集、存储和分析大数据，互联网公司尝试研发大数据技术，在众多技术方案中，开源系统 Hadoop 与 Spark 成为应用最广泛的大数据技术，由于它们的用户量巨大，已经初步成为大数据技术规范。

1.1.2　常见大数据应用场景

目前大数据技术被广泛应用在各个领域，它产生于互联网领域，并逐步推广到电信、医疗、金融、交通等领域，大数据技术在众多行业中产生了实用价值。

1. 互联网领域

在互联网领域，大数据被广泛应用在三大场景中，分别是搜索引擎、推荐系统和广告系统。

- **搜索引擎**：搜索引擎能够帮助人们在大数据集上快速检索信息，已经成为一个跟人们生活息息相关的工具。本书中涉及的很多开源大数据技术正是源于谷歌，谷歌在自己的搜索引擎中广泛使用了大数据存储和分析系统，这些系统被谷歌以论文的形式发表出来，进而被互联网界模仿。
- **推荐系统**：推荐系统能够在用户没有明确目的的时候根据用户历史行为信息帮助他们发现感兴趣的新内容，已经被广泛应用于电子商务（比如亚马逊、京东等）、电影视频网站（比如爱奇艺、腾讯视频等）、新闻推荐（比如今日头条等）等系统中。亚马逊科学家 Greg Linden 称，亚马逊 20%（之后一篇博文称 35%）的销售来自于推荐算法。Netflix 在宣传资料中称，有 60% 的用户是通过推荐系统找到自己感兴趣的电影和视频的。
- **广告系统**：广告是互联网领域常见的盈利模式，也是一个典型的大数据应用。广告系统能够根据用户的历史行为信息及个人基本信息，为用户推荐最精准的广告。广告系统通常涉及广告库、日志库等数据，需采用大数据技术解决。

2. 电信领域

电信领域是继互联网领域之后，大数据应用的又一次成功尝试。电信运营商拥有多年的数据积累，拥有诸如用户基本信息、业务发展量等结构化数据，也会涉及文本、图片、音频等非结构化数据。从数据来源看，电信运营商的数据涉及移动语音、固定电话、固网接入和无线上网等业务，积累了公众客户、政企客户和家庭客户等相关信息，也能收集到

电子渠道、直销渠道等所有类型渠道的接触信息，这些逐步积累下来的数据，最终形成大数据。目前电信领域主要将大数据应用在以下几个方面[⊖]：

- 网络管理和优化，包括基础设施建设优化、网络运营管理和优化。
- 市场与精准营销，包括客户画像、关系链研究、精准营销、实时营销和个性化推荐。
- 客户关系管理，包括客服中心优化和客户生命周期管理。
- 企业运营管理，包括业务运营监控和经营分析。
- 数据商业化：数据对外商业化，单独盈利。

3. 医疗领域

医疗领域的数据量巨大，数据类型复杂。到2020年，医疗数据将增至35ZB，相当于2009年数据量的44倍。医疗数据包括影像数据、病历数据、检验检查结果、诊疗费用等在内的各种数据，合理利用这些数据可产生巨大的商业价值。大数据技术在医疗行业的应用将包含以下方向：临床数据对比、药品研发、临床决策支持、实时统计分析、基本药物临床应用分析、远程病人数据分析、人口统计学分析、新农合基金数据分析、就诊行为分析、新的服务模式等[⊖]。

4. 金融领域

银行拥有多年的数据积累，已经开始尝试通过大数据来驱动业务运营。银行大数据应用可以分为四大方面[⊜]：

- **客户画像应用：**客户画像应用主要分为个人客户画像和企业客户画像。个人客户画像包括人口统计学特征、消费能力、兴趣、风险偏好等；企业客户画像包括企业的生产、流通、运营、财务、销售、客户、相关产业链上下游等数据。
- **精准营销：**在客户画像的基础上银行可以有效地开展精准营销，银行可以根据客户的喜好进行服务或者银行产品的个性化推荐，如根据客户的年龄、资产规模、理财偏好等，对客户群进行精准定位，分析出其潜在的金融服务需求，进而有针对性地进行营销推广。
- **风险管控：**包括中小企业贷款风险评估和欺诈交易识别等手段，银行可以利用持卡人基本信息、卡基本信息、交易历史、客户历史行为模式、正在发生的行为模式（如转账）等，结合智能规则引擎（如从一个不经常出现的国家为一个特有用户转账或从一个不熟悉的位置进行在线交易）进行实时的交易反欺诈分析。
- **运营优化：**包括市场和渠道分析优化、产品和服务优化等，通过大数据，银行可以监控不同市场推广渠道尤其是网络渠道推广的质量，从而进行合作渠道的调整和优化；银行可以将客户行为转化为信息流，并从中分析客户的个性特征和风险偏好，更

⊖ 傅志华：《大数据在电信行业的应用》

⊜ 吴闻新，《丁华：大数据在医疗行业的应用》，IDF 2013

⊜ 傅志华：《大数据在金融行业的应用》